Are VideoQA Models Truly Multimodal?

Ishaan Singh Rawal, Shantanu Jaiswal, Basura Fernando, Cheston Tan
Centre for Frontier AI Research, A*STAR, Singapore

1. Summary

Problem: Do Video Question Answering (VideoQA) models learn to align the multimodal information within and between the text and the video modalities? Or do they achieve high performance through shortcuts?

Motivation: The biases that are (i) present in the dataset and (ii) learnt and leveraged by the models are called shortcuts [1]. While most of the interpretability methods either focus on the model-centric or dataset-centric biases, we need a combined dataset-model centric approach to disambiguate the contribution of shortcuts in the model's performance.

Figure 1. Example of temporal bias. Does answering this question requires understanding the temporal sequence of frames?

Contributions:

(1) QUadrant AveraginG (**QUAG**) to find the contribution of specific modality interactions in the model's performance

(2) Counterfactual in Language And VIdeo (**CLAVI**) as a diagnostic for penalizing shortcut learning

Conclusions:

(1) Models achieve high accuracy on standard benchmarks even when the multimodal interactions are impaired

(2) Many models that perform well on standard datasets have trivial performance on CLAVI.

Our results show that **many current VideoQA models are incapable of multimodal understanding and rely on biases and shortcuts for their high performance**.

2. QUAG: Ablating Modality Interactions

- Focus on the self-attention based fusion modules, in which the modality embeddings are concatenated
- The modality interactions in the attention matrices are segregated in distinct quadrants (left panel of Fig. 2).
- We prove that consistent row-wise averaging of a set of quadrants leads to ablation of the particular interactions. This is known as short-circuiting (SC) (See Fig. 2).

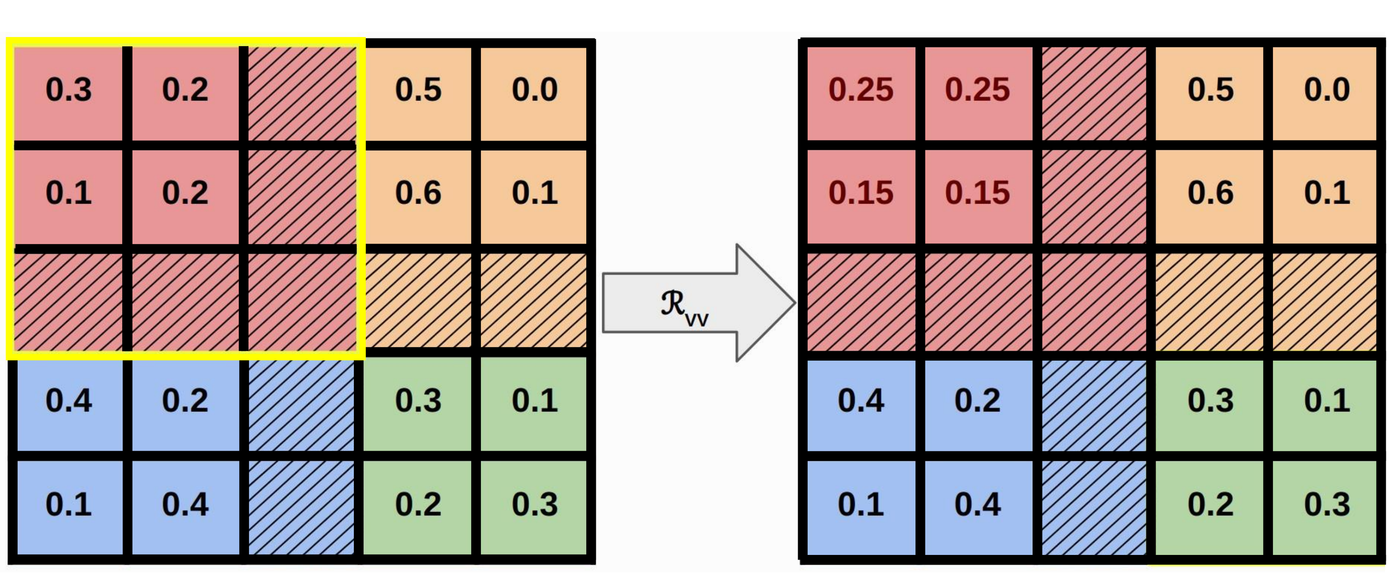

Figure 2. Toy example of row-wise averaging an attention matrix, with video embeddings pre-concatenated to the text. The quadrant colours denote the modality interactions (red: *VV*, yellow: *VT*, blue: *TV*, green: *TT*).

Table 1. % change in accuracy on SC (the name, based on the ablation effect in the first column) on ActivityNet-QA (A-QA) and NeXT-QA (N-QA)

Short-circuited quadrants	FrozenBiLM		JustAsk	
	A-QA	N-QA	A-QA	N-QA
{VV, TT} (unimodal)	-94.5%	-64.5%	-0.5%	-0.4%
{VT, TV} (crossmodal)	-25.9%	+0.7%	-1.0%	-0.6%
{TV, VV} (video)	-1.1%	+0.0%	-1.3%	-0.7%
{VT, TT} (text)	-96.8%	-63.3%	-0.3%	-0.2%

The results of QUAG are summarized in Table 1

1. The drop in video-SC is ~1%; hence, the models don't rely on core features of the video for their performance
2. FrozenBiLM relies on text (drops in unimodal and text SC) but leverages crossmodal interactions for A-QA only
3. JustAsk does not rely on core multimodal features (insignificant drop for all the SC operations)

This means that **high performance on standard datasets does not imply joint multimodal understanding.**

3. CLAVI: Counterfactual Diagnosis

Temporal understanding requires aligning both video and text [2]. We curate temporal questions (***before/after*** or ***beginning/end***) from annotations with counterfactuals in:

1. **Language**: By replacing *before/after* or *beginning/end*
2. **Video**: By swapping the order of frames (Refer to Fig. 3)

We also add **existence** questions (*Does <event> occur?*), and **negative control** questions containing events that do not occur within video for benchmarking shortcut learning.

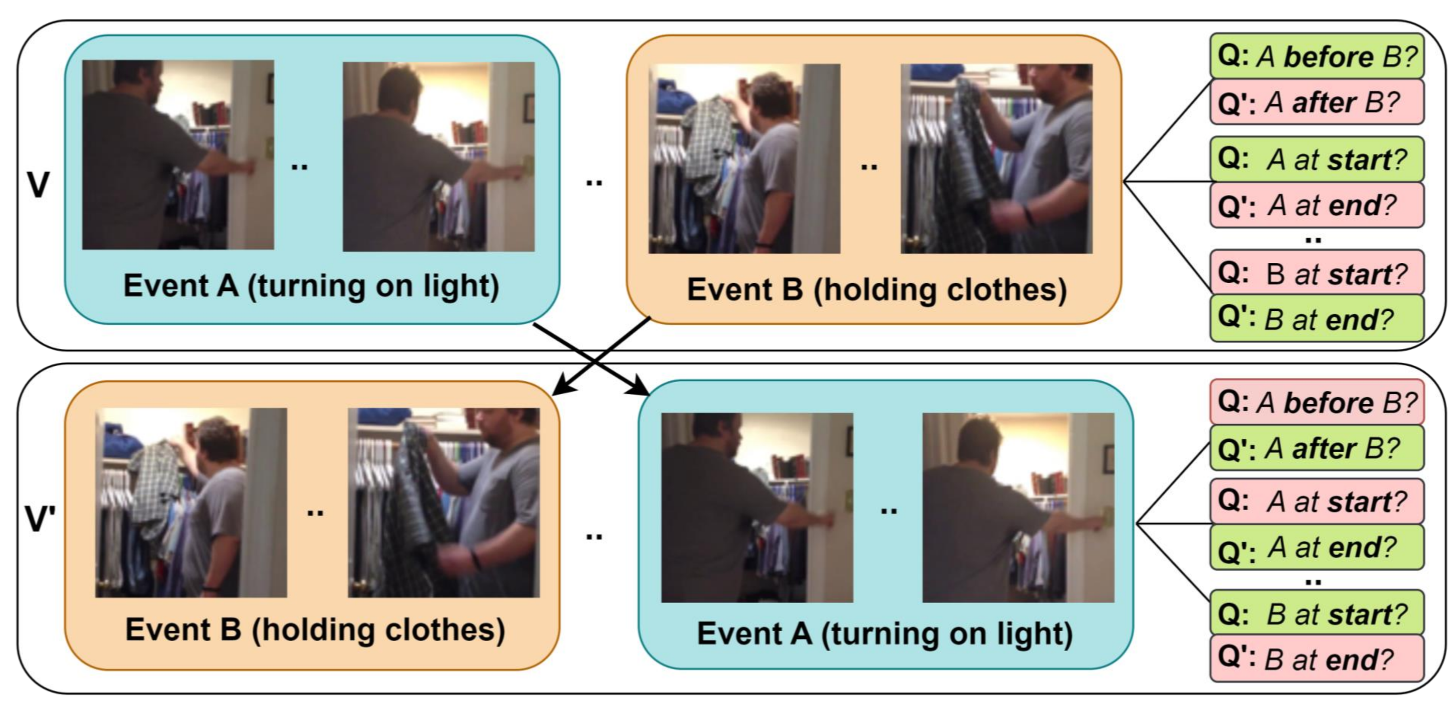

Figure 3. Curation of video and text counterfactuals (V', Q'), from video and question (V, Q). The colour represents the answer (red: no, green: yes)

We report consistent accuracies as:

$$Cacc_V = \mathbb{1}\{ F(V, Q) == A_{VQ} \text{ AND } F(V', Q) == A_{V'Q} \}$$
$$Cacc_T = \mathbb{1}\{ F(V, Q) == A_{VQ} \text{ AND } F(V, Q') == A_{VQ'} \}$$

where F is the model and A_{VQ} is the answer of the input (V, Q)

Table 2. Finetuning results on CLAVI. Note that the existence and negative control form the control subset and the rest, counter(factual)

Subset	Metric	JustAsk	FrozenBiLM	Singularty-T	All-in-one
Control	$Cacc_V$	98.0	93.2	92.7	98.1
	$Cacc_T$	98.2	93.7	93.5	98.2
Counter	$Cacc_V$	3.6	54.1	1.7	1.2
	$Cacc_T$	2.4	57.2	0.5	0.8

Finetuning Results on CLAVI (Table 2):

1. Models achieve >90% score for the control subset that does not require joint multimodal understanding
2. Except FrozenBiLM, models achieve <3% performance on the counter subset that penalizes shortcut learning

This reveals that **many models that achieve high accuracy on benchmarks are incapable of joint multimodal understanding, creating its illusion through shortcuts**

Questions? Feedback?
Please scan the QR code or email
rawal_ishaan_singh@cfar.a-star.edu.sg

References:
[1] Murali, Nihal, et al. "Beyond Distribution Shift: Spurious Features Through the Lens of Training Dynamics." *TMLR* (2023)
[2] Bagad, Piyush, et al. "Test of Time: Instilling Video-Language Models with a Sense of Time." *CVPR (2023)*